A Parsed Linguistic Atlas of Early Middle English

A Parsed Linguistic Atlas of Early Middle English

Rob Truswell, Rhona Alcorn, and James Donaldson
University of Edinburgh
rob.truswell@ed.ac.uk, r.alcorn@ed.ac.uk,
james.donaldson@ed.c.uk

AMC symposium, 10/6/16

Outline

- An initial report on a new, BA/Leverhulme-funded, corpus project.
- Goal: use data from the Linguistic Atlas of Early Middle English to plug a gap in the Penn Parsed Corpora of Historical English.
- Agenda:
  1. Beginner's guide to PPCHE
  2. Beginner's guide to LAEME
  3. Plans for the project
  4. Beyond this project

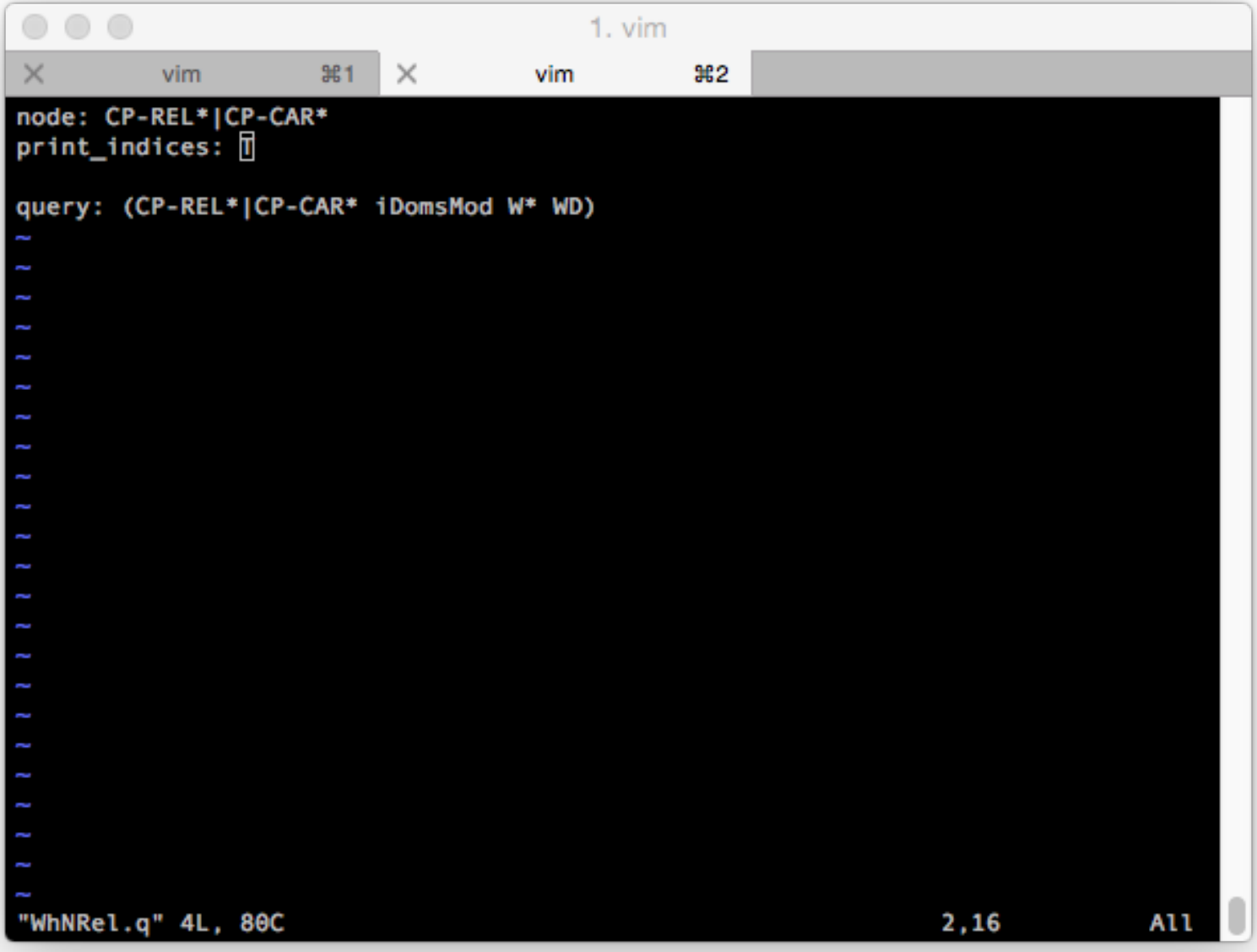
Section 1

PPCHE

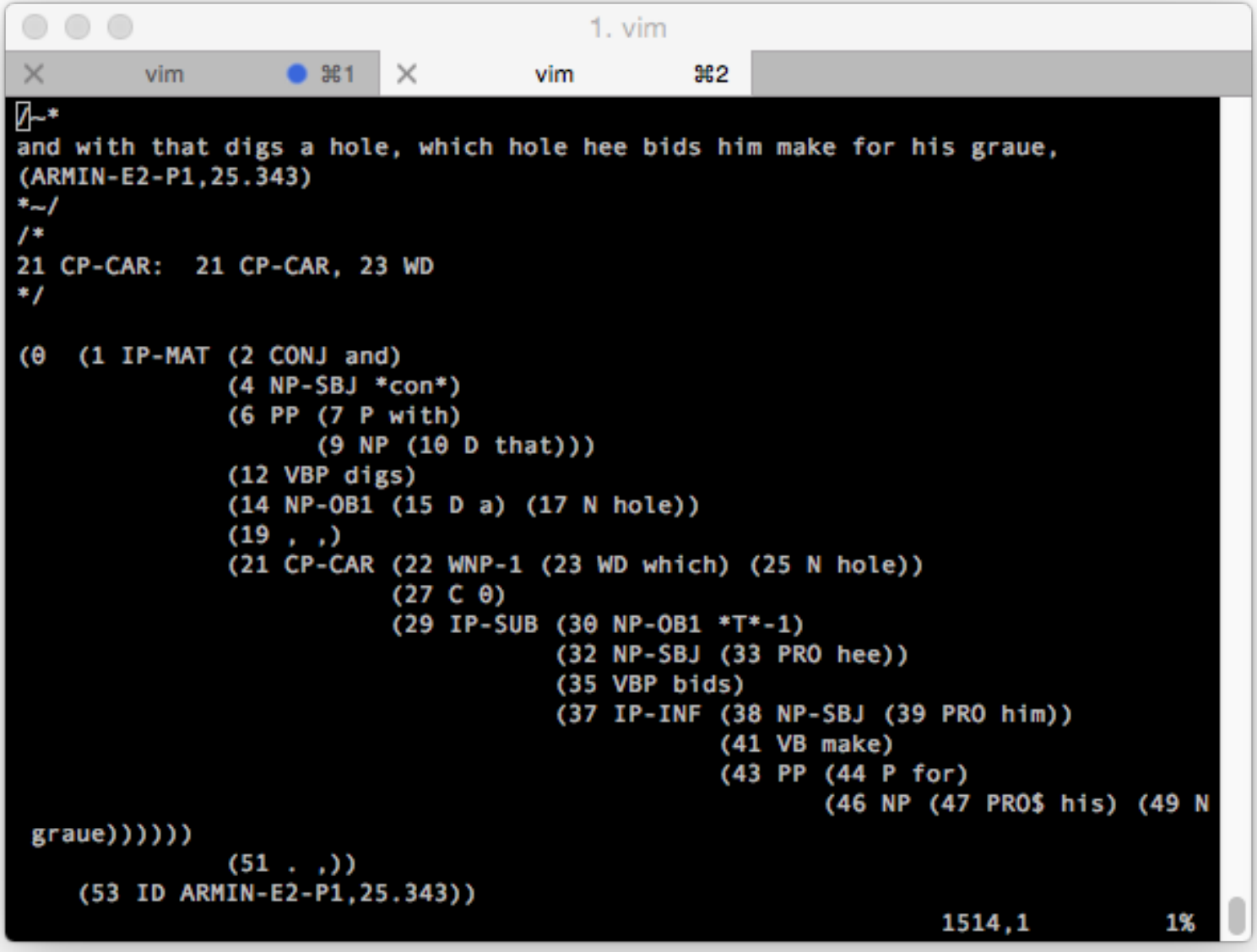
The Penn historical corpus format

- Family of parsed historical corpora covering English (7.7m words, –1914) and several other Germanic and Romance languages.
- Annotated with detailed labelled brackets.
  - Explicit annotation of grammatical function.
  - Explicit representation of many cases of "movement", extraposition, etc.
- The PPCHE philosophy:
  - no claim to being an accurate theory of syntax.
  - aim to include useful information about constituency in a way that can be consistently implemented.
  - So (almost) no VP nodes, default high attachment, etc.
- Major virtue: very easy to query information about phrasal syntax.
- Very easy to make (some kind of) crosslinguistic comparison.

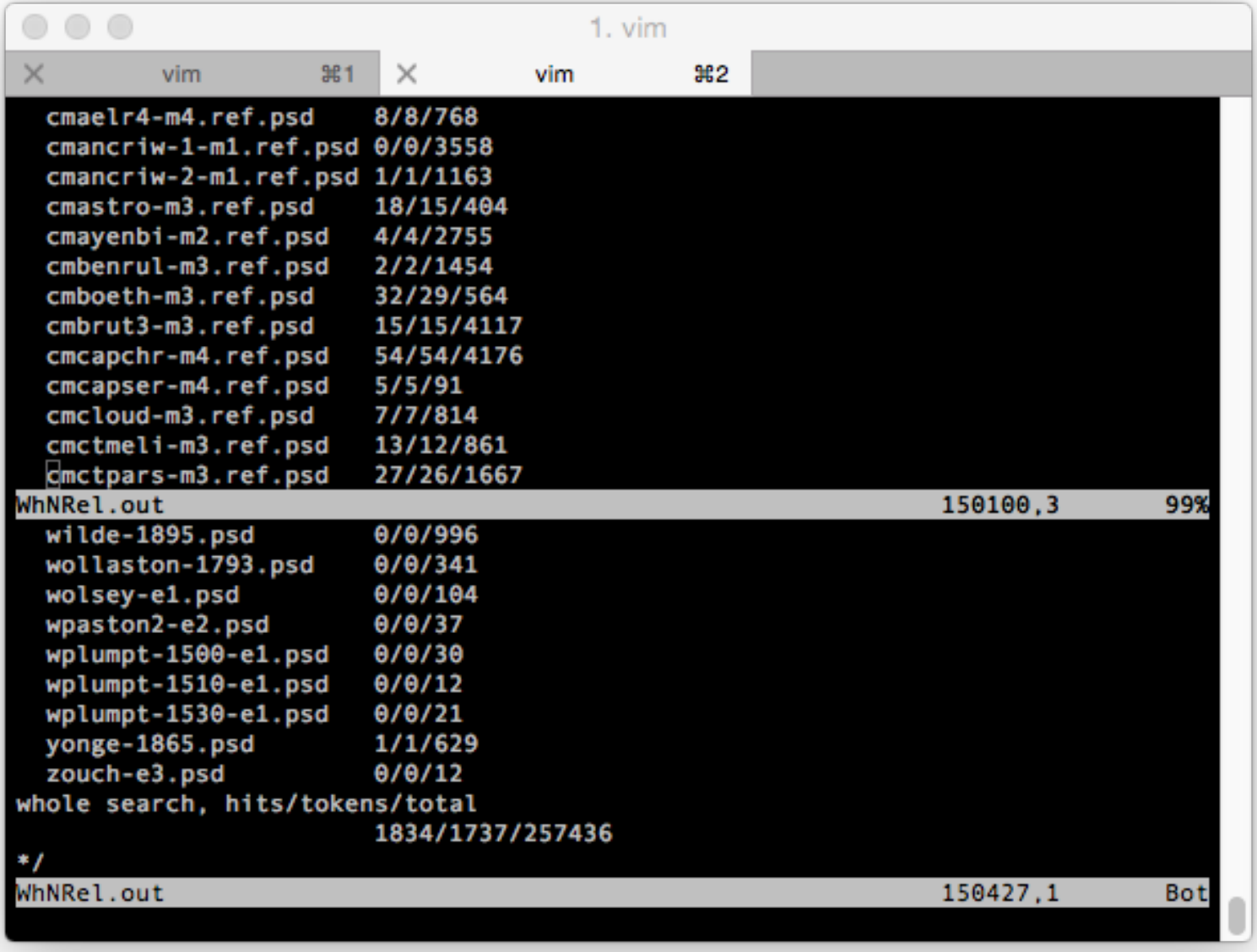
Sample query



Sample query: sample output



Sample query: counts



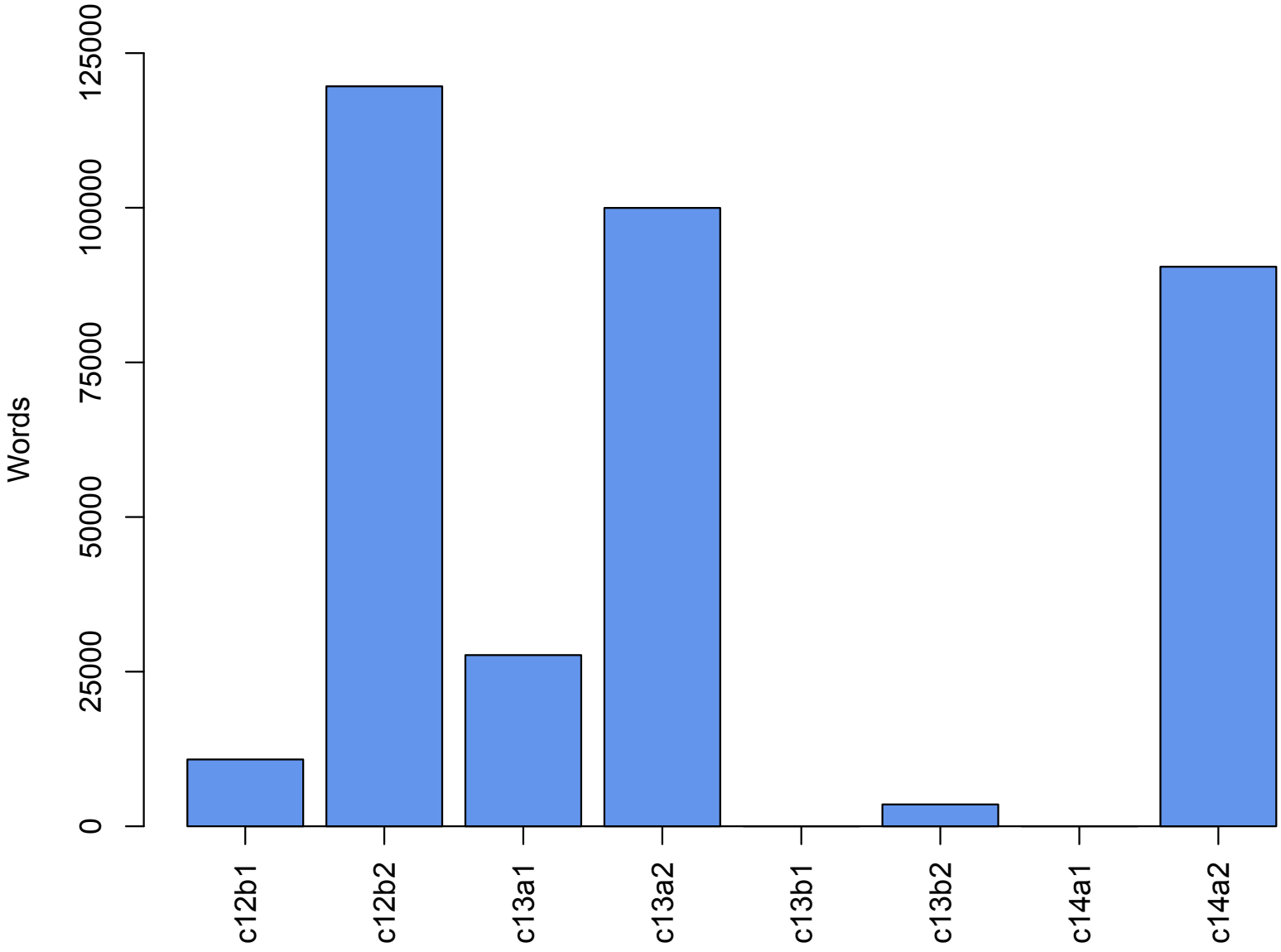
PPCME2 weaknesses

- Built from published editions, not manuscripts.
- Limited metadata.
- Not lemmatized.
- Significant data gap c.1250–1340.

The data gap: PPCME2, 1150–1350

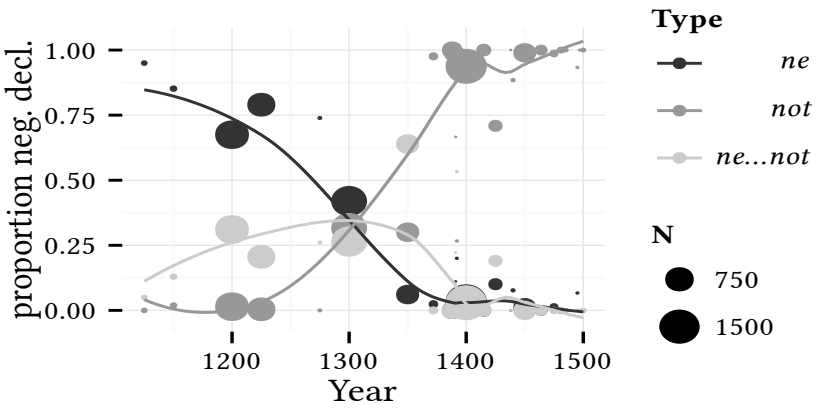
Filename	Title	Date	Words
cmkentho	Kentish Homilies	c12a2–b1	4048
cmpeterb	Peterborough Chronicle	c.1131, c.1154	6757
cmlambx1	Lambeth Homilies	c12b2	20752
cmtrinit	Trinity Homilies	c12b2	41844
cmorm	Ormulum	c12b2	50579
cmlamb1	Lambeth Homilies	c12b2	6459
cmvices1	Vices and Virtues	c13a1	27677
cmsawles	Sawles Warde	c13a2	4111
cmhali	Hali Meiðhad	c13a2	8495
cmkathe	St. Katherine	c13a2	8699
cmjulia	St. Juliana	c13a2	6810
cmmarga	St. Margaret	c13a2	8069
cmancriw	Ancrene Riwle	c13a2	63790
cmkentse	Kentish Sermons	c13b2?	3534
cmayenbi	Ayenbite of Inwyt	1340	45944
cmearlps	Earliest Prose Psalter	c.1350	44521

The data gap: PPCME2, 1150–1350

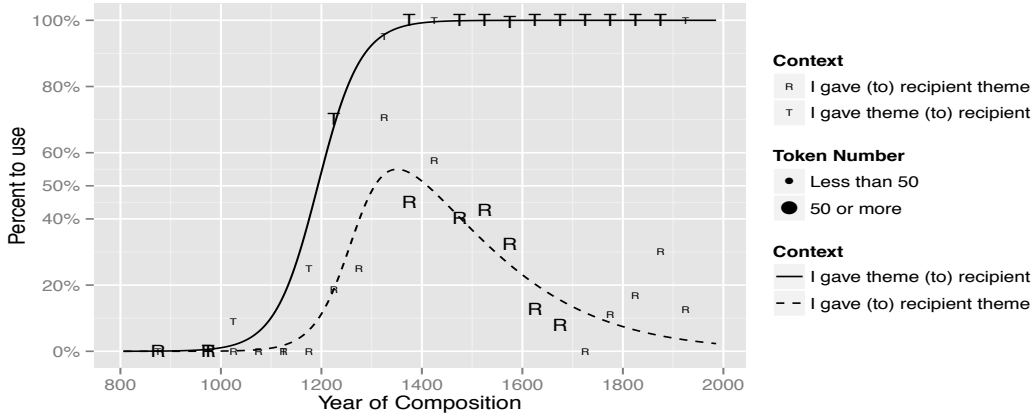


The data gap in action

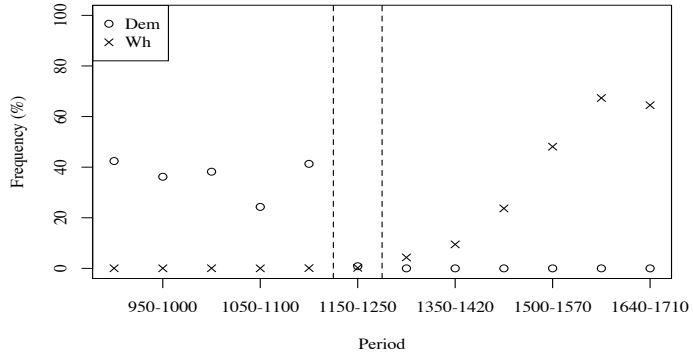
Ecay & Tamminga (2016): Expression of negation



Bacovcin (2016): Double object word order



Gisborne & Truswell (2016): Argument relatives



Help is at hand

- Luckily, the Linguistic Atlas of Early Middle English (Laing 2013) complements PPCME2 remarkably well.
  - Built from manuscripts.
  - Extensive metadata.
  - Lemmatized.
  - Significant amounts of material from 1250–1325.
- A major motivation for constructing PLAEME is that LAEME complements PPCME2 so well.
- By transferring features from LAEME, we also have the opportunity to expand on the PPCHE format.

Section 2

LAEME

LAEME

- Compilation began early '90s; published online 2008
- Specimens of written English 1150 – 1325:
  - Documents, e.g. Chertsey Cartulary
  - Prose, e.g. *Vices and Virtues*
  - Poetry, e.g. *Poema Morale*
  - Lyrics, e.g. *Love Song of Our Lady*
- c.650,000 words (shortest 18; longest 30,237)

Transcription

- Upper case for 'plain text' ms letters
- * for capitals
- Lower case for special characters:
  - y = þ, d = ð, w = ƿ, ae = æ, g = ȝ, z = ʒ

MS: <For it is caldore ner þe se> (*The Infancy of Christ*)

LAEME: *FOR IT IS CALDORE NER yE SE

Tagging

- Most morphemes (incl. inflections & affixes) are tagged
- $lexel/grammel_form, e.g. $for/cj_*FOR
  - lexel: semantic identifier
  - grammel: grammatical identifier
- All tags can be translated into lemmas
- Word-level tags arranged vertically in tagged text files

```
$for/cj_*FOR
$/P13NI_IT
$be/vps13_IS
$cold/aj-cpv_CALD+ORE $-er/xs-cpv-aj_+ORE
$near/pr_NER
$/T<pr_yE
$sea/n<pr_SE
```

Syntactic clues: NPs

```
$near/pr_NER
$/T<pr_yE
$sea/n<pr_SE
```

- $/T(…) = Definite determiner
- $xxx/n(…) = Noun
- $/T(…) + $xxx/n(…) = NP

Syntactic clues: NP function

```
$near/pr_NER
$/T<pr_yE
$sea/n<pr_SE
```

- <pr = ‘in the scope of a preceding preposition’
- $/T<pr + $xxx/n<pr = NP object of P

Syntactic clues: NP function

<To teche þe volk þe riʒte lawe>
'to teach the folk the right law'

```
$to/im+C_*TO
$teach/vi-m_TECH+E $/vi-m_+E
$/TOi_yE
$folk/nOi_VOLK
$/TOd_yE
$right/ajOd_RIzTE
$law/nOd_LAWE
```

- $/T(…) + $folk/n(…) = NP
- Oi in grammel indicates indirect object function

Syntactic clues: NP function

```
$to/im+C_*TO
$teach/vi-m_TECH+E $/vi-m_+E
$/TOi_yE
$folk/nOi_VOLK
$/TOd_yE
$right/ajOd_RIzTE
$law/nOd_LAWE
```

- $/T(…) + $right/aj(…) + $law/n(…) = NP
- Od in grammel indicates direct object function

Non-local dependency: discontinuous XPs

<... ðat ghe ne migte **hi*m*** bringen **on**>
'... which she may not prove **against him**'

- >pr = ‘object of a following preposition’
- pr< = ‘preposition with a preposed object’

```
$/RTIOd_dAT
$/P13NF_GHE
$/neg-v_NE
$may/vpt13_MIGTE
$/P13>prM_HIm
$bring/vi_BRING+EN $/vi_+EN
$on{p}/pr<{rh}_ON
```

Non-local dependency: 'braced' negation

- >= points forward to a non-contiguous coordinating item
- <= points back to a non-contiguous coordinating item

<þou **ne** sselt **naʒt** *con*senti ...>
'thou shall not consent ...'

```
$/P12N_yOU
$/neg-v>=_NE
$shall/vps12_SSELT
$not/neg-v<=_NAzT
$consent/viFir_conSENT+I $/viFir_+I
```

LAEME: Text selection

Prioritised texts:

1. 1250 - 1325
2. > 100 words
3. No pre-existing parsed version

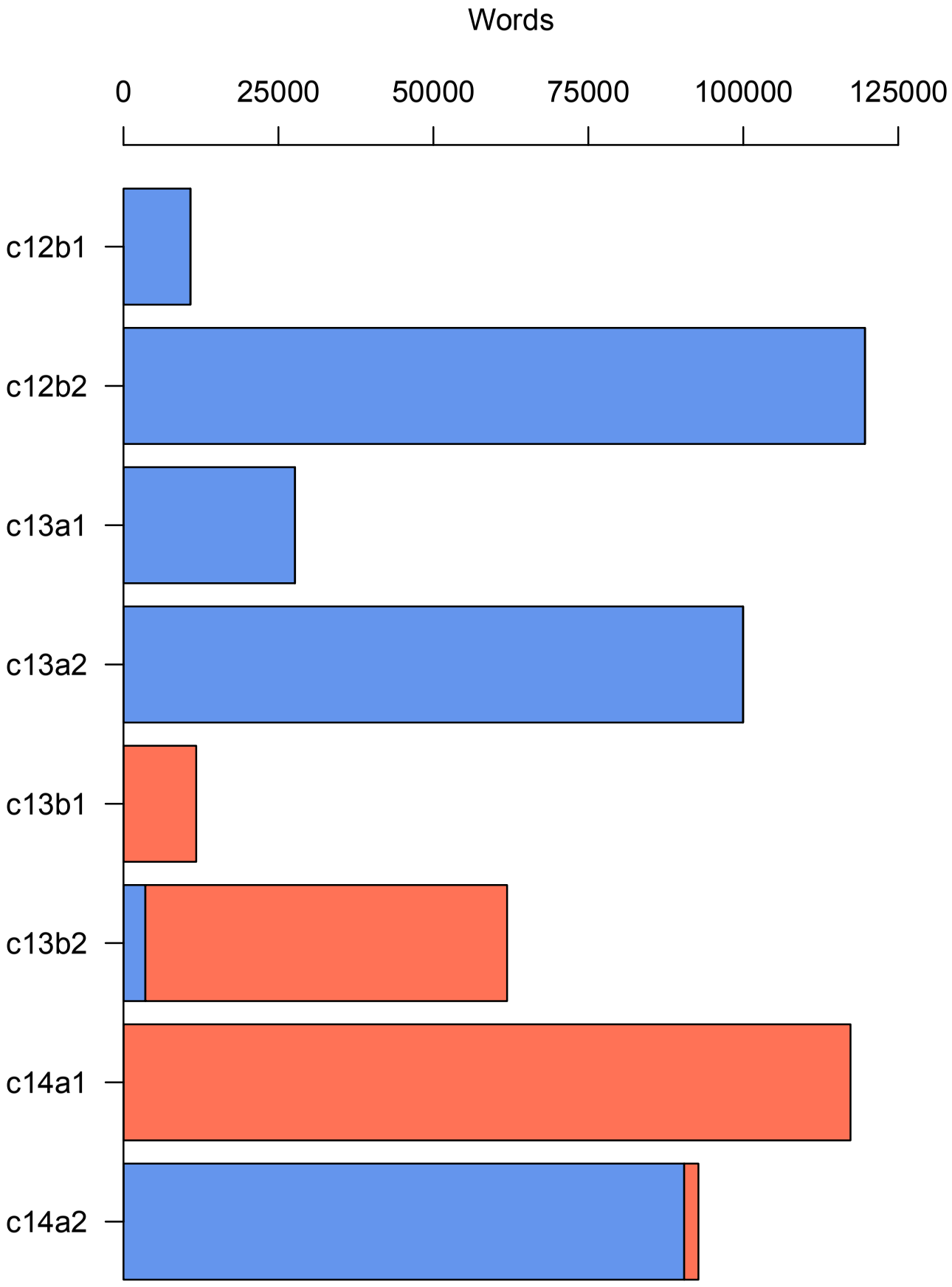
Multiple versions of same text:

4. Choose on basis of date, length and county

Our top 8

Text	#	Date	Words	County
S. English Legendary	286	ca.1310-20	30,237	Berks
Nthn Homily Coll'n	298	C14a	22,164	Yorks NR
Havelok	285	C14a1	17,089	Norfolk
Cursor Mundi, Hand A	297	C14a	15,107	Yorks ER
Cursor Mundi, Hand C	296	C14a	14,087	York
Infancy of Christ	283	ca.1300	12,489	Oxon
Genesis and Exodus	155	C14a1	12,467	Norfolk
Life of Christ	281	ca.1300	10,547	Oxon

These texts alone supplement PPCHE by 134,187 words


Words
0
25000
50000
75000
100000
125000
c12b1
c12b2
c13a1
c13a2
c13b1
c13b2
c14a1
c14a2

Section 3

The parsing process

Distinctive properties

- Unusual parsing approach:
  - LAEME grammels stand in a many–many relationship with PPCHE POS tags, so will not be directly visible in the output.
  - Nevertheless, LAEME grammels contain a large amount of syntactically relevant information:
    - Category
    - Grammatical function
    - Some meaning relations
    - Some nonlocal dependencies

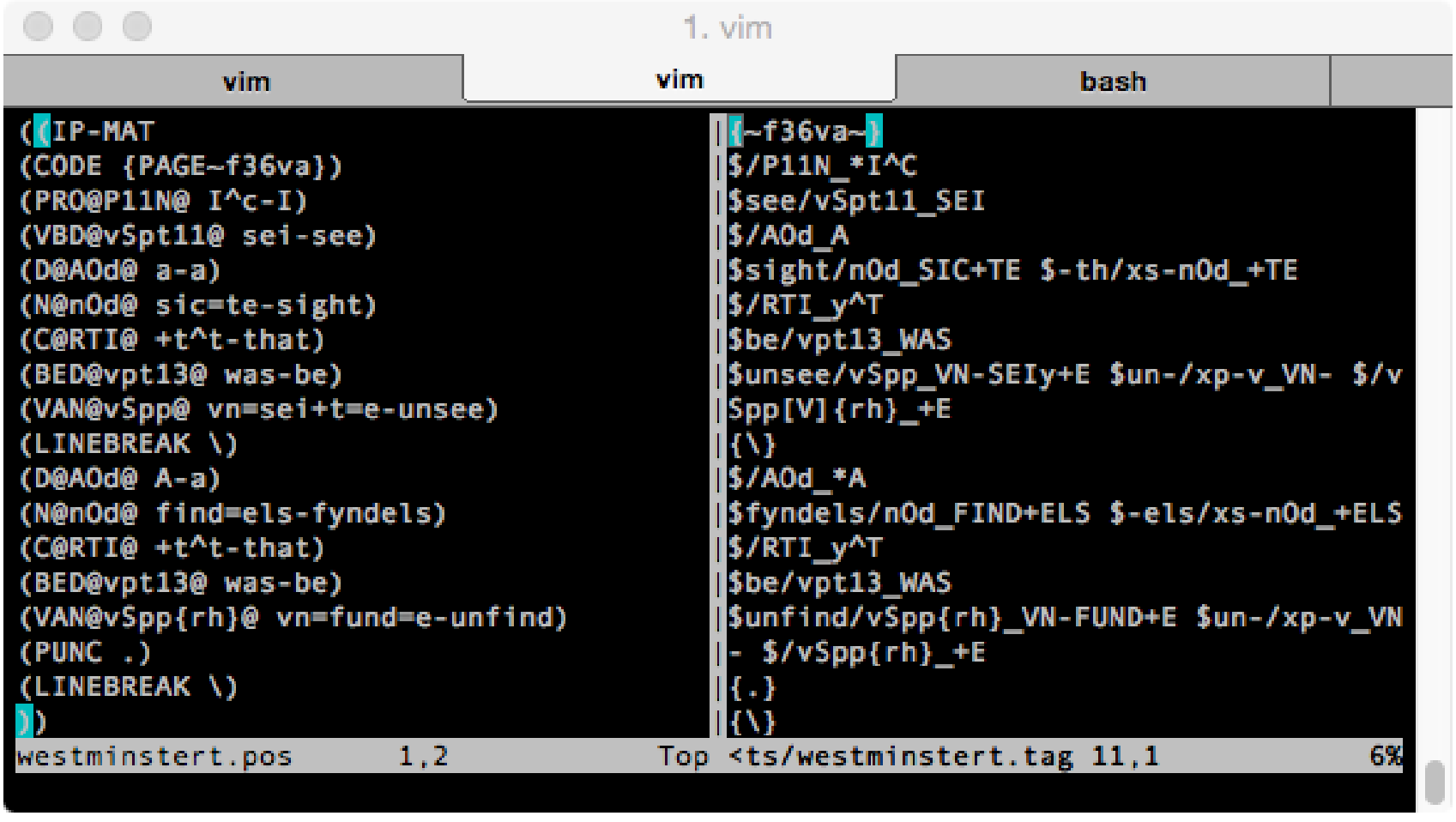
      `$/neg-v([neither]<av>norC>nor>norC)_NE`

      (laud108bt.tag, line 1161)
  - Plan in essence: project the LAEME grammels into labelled brackets, while also replacing them with PPCHE tags.
  - Allows for fairly accurate and very detailed automatic structure generation.

Workflow: Format conversion

1. Take LAEME text
2. Store initial metadata as separate file.
3. For lines with textual material (initial character $, ’, ;):
   - 3.1 Find full word form, most informative grammel, etc.
   - 3.2 Segment that material into `$lexel/grammel_form`.
   - 3.3 Convert LAEME orthography into PPCHE orthography.
   - 3.4 Add lemmata for function words (no lexels in LAEME).
   - 3.5 Reformat as `(POS@grammel@ form-lexel)`, where POS is a best-guess PPCHE equivalent of the LAEME grammel, with some lexel-by-lexel correction for e.g. different treatment of relativizers, subordinating conjunctions, etc.).
   - 3.6 Split compound forms (*ifthat, nolde, shalþu*, etc.), retag, etc.
4. For other nodes (detailing deletion, insertion, commentary, etc., also punctuation), figure out an appropriate tag (e.g. `COMMENT`, `INS`), reformat slightly, and pass to output file.
5. Use punctuation as (*very*) rough guide to sentence boundaries, insert `IP-MAT` brackets accordingly.

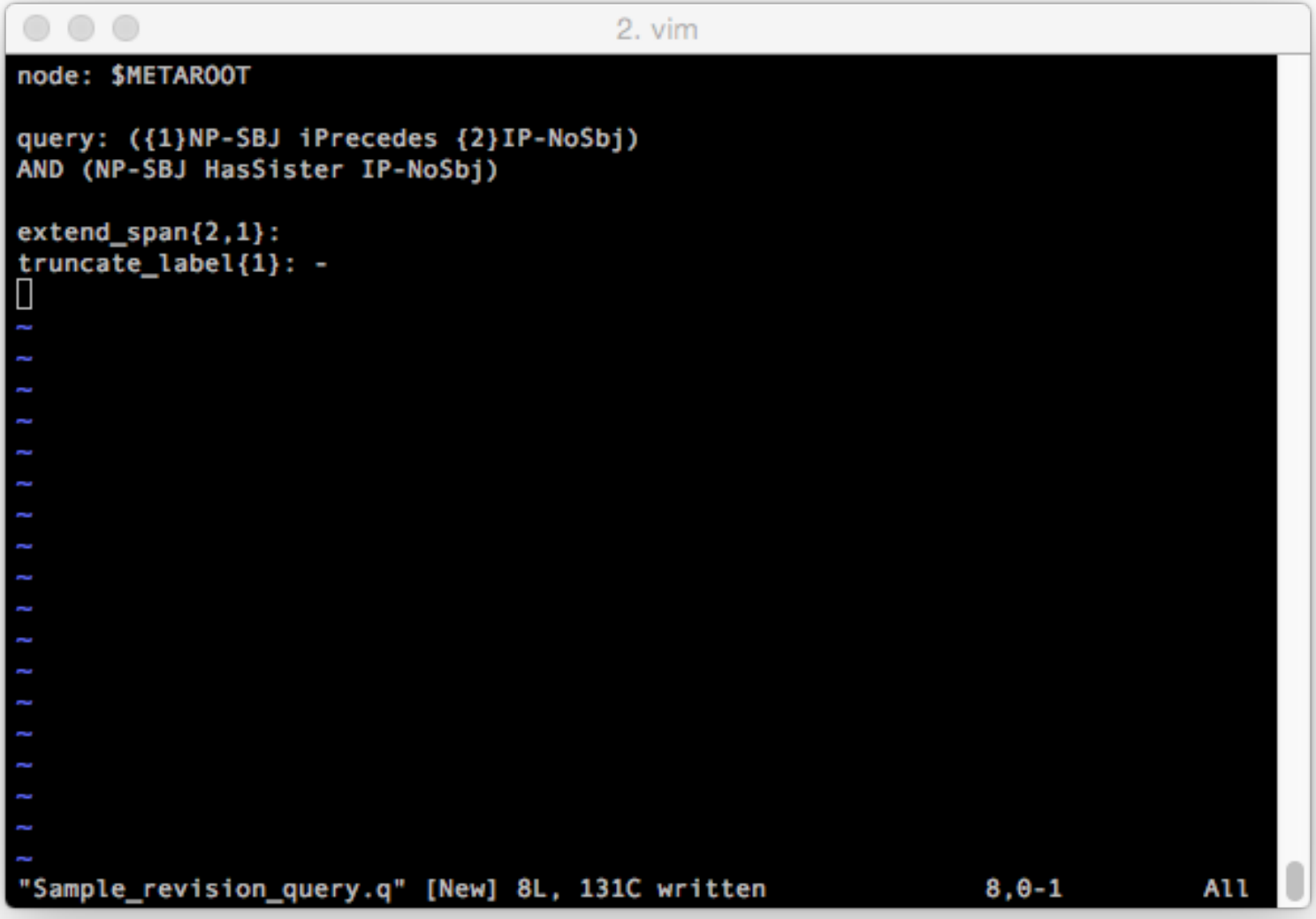
So far


1. vim
vim
vim
bash
((IP-MAT
(CODE {PAGE~f36va})
(PRO@P11N@ I^c-I)
(VBD@vSpt11@ sei-see)
(D@AOd@ a-a)
(N@nOd@ sic=te-sight)
(C@RTI@ +t^t-that)
(BED@vpt13@ was-be)
(VAN@vSpp@ vn=sei+t=e-unsee)
(LINEBREAK \)
(D@AOd@ A-a)
(N@nOd@ find=els-fyndels)
(C@RTI@ +t^t-that)
(BED@vpt13@ was-be)
(VAN@vSpp{rh}@ vn=fund=e-unfind)
(PUNC .)
(LINEBREAK \)
))
{~f36va~}
$/P11N_*I^C
$see/vSpt11_SEI
$/AOd_A
$sight/nOd_SIC+TE $-th/xs-nOd_+TE
$/RTI_y^T
$be/vpt13_WAS
$unsee/vSpp_VN-SEIy+E $un-/xp-v_VN- $/v
Spp[V]{rh}_+E
{\}
$/AOd_*A
$fyndels/nOd_FIND+ELS $-els/xs-nOd_+ELS
$/RTI_y^T
$be/vpt13_WAS
$unfind/vSpp{rh}_VN-FUND+E $un-/xp-v_VN
- $/vSpp{rh}_+E
{.}
{\}
westminstert.pos 1,2 Top
<ts/westminstert.tag 11,1 6%

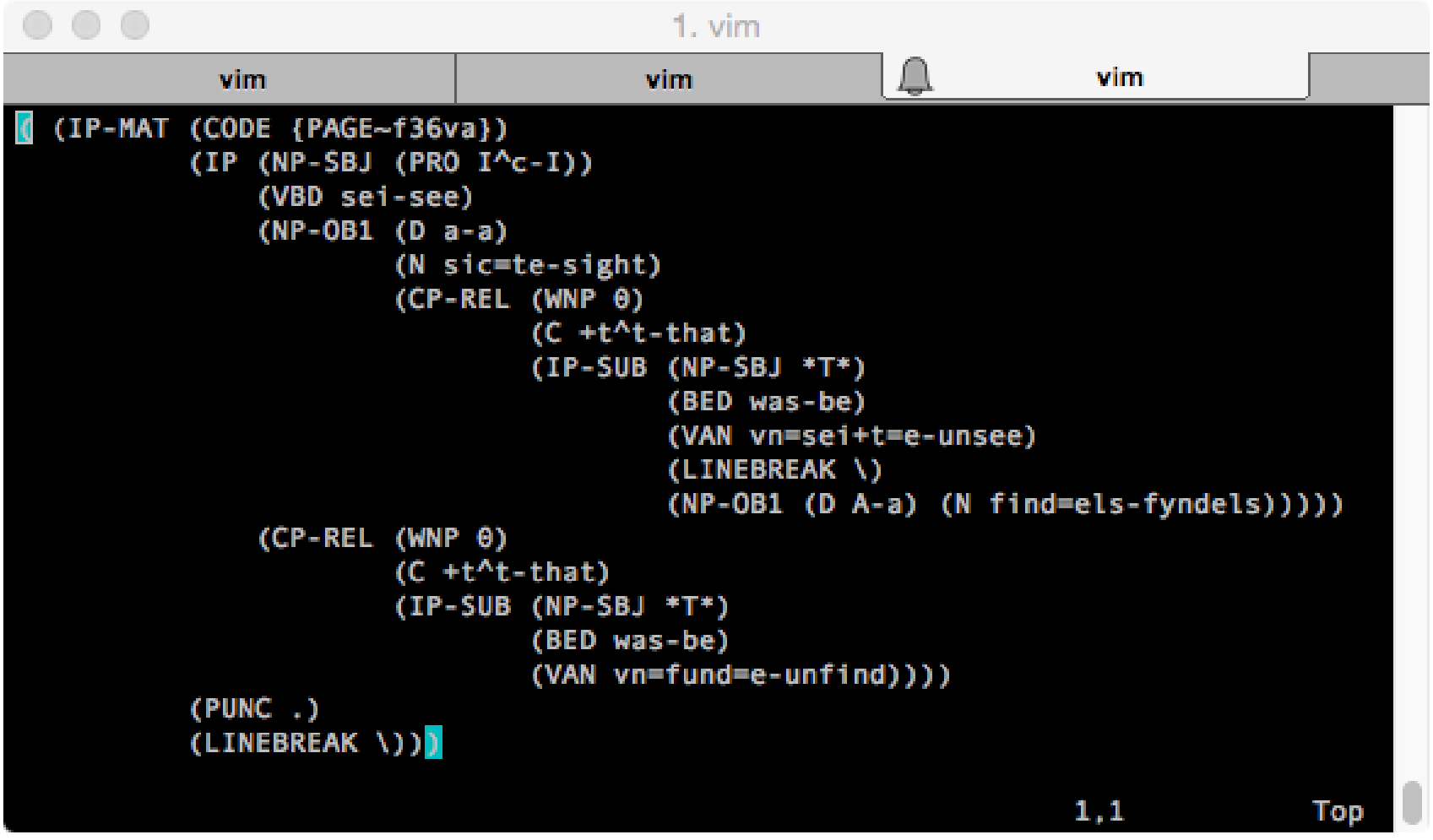
Workflow: Parsing

1. Use corpus revision queries to incrementally add/adjust bracketing, relabel, etc.
2. Transfer rhyme information to word-level tags (not yet implemented).
3. Strip out LAEME grammels.
4. Manually correct automatically generated parse.

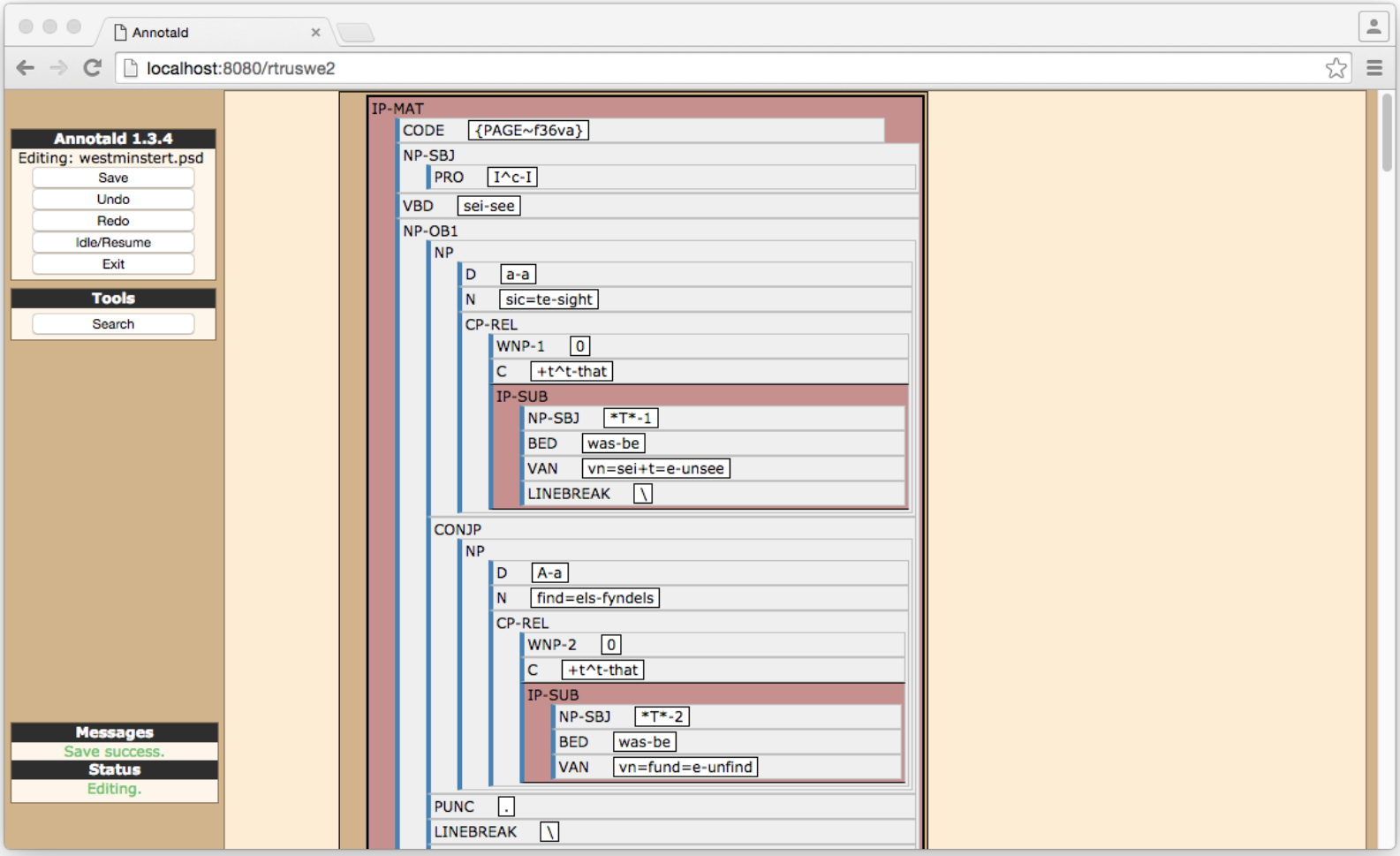
Sample corpus revision query



Automatically generated output



Hand-correction: Annotald



Section 4

What next?

Prospects

- Initial grant covers $< \frac{1}{3}$ of LAEME.
- Longer-term goals: the rest of LAEME, and LAOS.
- Practical issues:
  - Up-to-speed annotators can correct 4-500 words/hour, so working through the rest of LAEME and LAOS involves c.1500 hours of someone's time (Jim's?).
  - Any way to "reuse parses" across parallel texts? Rough calculation: c.390k words of LAEME are versions of texts already parsed in PPCME2/PCMEP or slated for inclusion in our first sample ($\approx 87\%$ of the remainder of LAEME).
- Potential for new research uses of parsed historical corpora?
  - Traditionally, PPCHE people have had limited interest in dialectal variation.
  - Our initial choice of texts for parsing was motivated by helping PPCHE people to find fuller answers to PPCHE-style questions.
  - Any hope for using parallel texts in a fully parsed version of LAEME to investigate dialectal variation in syntactic structure?

References

Bacovcin, H. A. (2016). Modelling interactions between morphosyntactic changes. In E. Mathieu & R. Truswell (Eds.), *From Micro-change to Macro-change*. Oxford: Oxford University Press.

Ecay, A. & Tamminga, M. (2016). Persistence as a diagnostic of grammatical status: The case of Middle English negation. In E. Mathieu & R. Truswell (Eds.), *From Micro-change to Macro-change*. Oxford: Oxford University Press.

Gisborne, N. & Truswell, R. (2016). Where do relative specifiers come from? In E. Mathieu & R. Truswell (Eds.), *From Micro-change to Macro-change*. Oxford: Oxford University Press.

Laing, M. (2013). A Linguistic Atlas of Early Middle English, 1150–1325. Version 3.2, `http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html`.